Title

Stronger instruments and refined covariate balance in an observational study of the effectiveness of prompt admission to intensive care units

Journal

Journal of the Royal Statistical Society. Series A: Statistics in Society, 183(4)

ISSN

0964-1998

Authors

Keele, L
Harris, S
Pimentel, SD
et al.

Publication Date

2020-10-01

DOI

10.1111/rssa.12437

Peer reviewed

Stronger Instruments and Refined Covariate Balance in an Observational Study of the Effectiveness of Prompt Admission to the ICU*

Luke Keele† Steve Harris ‡ Samuel D. Pimentel§ Richard Grieve¶

June 7, 2018

Abstract

Instrumental Variable (IV) methods, subject to appropriate identification assumptions, allow for consistent estimation of causal effects in the presence of unobserved confounding. Near-far matching has been proposed as one analytic method to improve inference by strengthening the effect of the instrument on the exposure and balancing observable characteristics between groups of subjects with low and high values of the instrument. However, in settings with hierarchical data (e.g. patients nested within hospitals), or where several covariate interactions must be balanced, conventional near-far matching algorithms may fail to achieve the requisite covariate balance. We develop a new matching algorithm, that combines near-far matching with refined covariate balance, to balance large numbers of nominal covariates while also strengthening the IV. This extension of near-far matching is motivated by a case study that aims to identify the causal effect of prompt admission to the Intensive Care Unit on 7-day and 28-day mortality.

*We thank Anirban Basu for comments and discussion.

†Professor, McCourt School of Public Policy, Georgetown University, 37th & O St, NW Washington DC 20057 Email: luke.keele@gmail.com, corresponding author.

‡Clinical Lecturer in Anaesthesia and Critical Care, University College London, Hospital, Email: doc@steveharris.me

§Department of Statistics, University of Pennsylvania, Philadelphia, PA, Email: spi@wharton.upenn.edu

¶Professor of Health Economics Methodology, Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, 15-17 Tavistock Place, London, WC1H 9SH, Email: richard.grieve@lshtm.ac.uk

1 Introduction

Instrumental variable methods can address the problem of unobserved confounding in observational studies. However, a common concern is that the instrument is weak and liable to provide biased estimates if there are deviations from random assignment of the instrument (Small and Rosenbaum 2008). Near-far matching addresses this concern by creating pairs that are similar in terms of observed covariates but dissimilar on the values of the instrument (Baiocchi et al. 2010). However, the matching algorithms on which near-far matching is based may be unable to balance many-valued nominal covariates (Pimentel et al. 2015). Such nominal covariates may be present due to hierarchical data structures, for example pupils nested within schools, or patients within hospitals. Alternatively, the interaction of several nominal covariates produces nominal variables of this type. Here, we extend extant near-far matching algorithms based on optimal nonbipartiate and integer programming (Baiocchi et al. 2010; Zubizarreta et al. 2013), to include refined covariate balance, which is designed to optimally balance many-valued nominal covariates. (Pimentel et al. 2015). Our approach is novel in that we develop a single matching algorithm that combines refined covariate balancing with near-far matching to handle the more complex data structures found in some observational studies. Specifically the use of refined covariate balance allows us to match within many blocks and to balance nominal covariates and their higher order terms.

Our methodological contribution is motivated by an empirical investigation of whether prompt admission to the intensive care unit (ICU) reduces mortality. Previous studies have reported that ICU admission is associated with higher mortality, but these studies are likely to have provided biased estimates, as patients promptly admitted to the ICU tend to have a more severe case-mix, according to both measured and unmeasured patient characteristics (Chalfin et al. 2007; Gabler et al. 2013; Renaud et al. 2009). To address this problem, other investigators have used an instrumental variable to study the effects of critical care (Hu et al. 2018; Kc and Terwiesch 2012; Shmueli et al. 2004; Valley et al. 2015; Pirracchio et al. 2011). An instrumental variable acts as a nudge or encouragement for treatment receipt that can only affect the outcome via the treatment.

Specifically, our study uses the level of bed-occupancy at the time of patient assessment as an instrumental variable for the effect of prompt ICU admission on mortality (Harris et al. 2018). We conduct near-far matching to strengthen the instrument, while recognizing that unobserved confounding may be correlated within hospitals. Our matching algorithm therefore incorporates refined balance constraints and optimal subsetting of the data to conduct the near-far matching, while fine-balancing patients within hospitals and the interaction of several nominal covariates. Thus our key contribution is to account for hospital specific components in the near-far match.

The paper proceeds as follows. In Section 2, we outline the empirical investigation and provide an overview of the assumptions required for IV identification in the context of this example. Section 3 provides details on near-far matching including the extension with refined covariate balance. Section 4 details the implementation, and Sections 5 and 6 report results. Finally, in Section 7, we conclude.

2 Motivating Example, Notation, and Causal Framework

We assess the effect of prompt ICU admission as part of a prospective cohort study of patients with deteriorating health referred for assessment for ICU admission in 48 UK National Health Service (NHS) hospitals between 1 November 2010 and 31 December 2011 (the SPOTlight study) (Harris et al. 2015). The patients included were those on general hospital wards that were considered for ICU admission. However, due to logistical and temporal constraints on ICU bed availability, only a subsample of these patients were transferred promptly to an ICU. The study uses this variation to define the exposure (treatment) as transfer to the ICU within four hours of assessment. The non-exposed (control) group is defined as those patients who were not admitted to ICU or admitted with a delay of more than four hours. The four hour cutoff was chosen a priori according to published clinical guidelines for the UK (The Intensive Care Society 2013). The main endpoints of interest were 7- and 28-day mortality.

The level of ICU bed occupancy at the time of assessment for ICU admission was recorded

for each patient, and we use it as an instrument for prompt ICU admission. Other covariates recorded at assessment include age, sepsis (0/1), peri-arrest (0/1), physiological measures such as the Intensive Care National Audit & Research Centre (ICNARC) physiology score, the NHS National Early Warning Score (NEWS), and the Sequential Organ Failure Assessment (SOFA) score. The level of care at assessment, and recommended level of care after assessment were defined using the UK Critical Care Minimum Dataset (CCMDS) levels of care (0 and 1 for normal ward care, 2 for care within a high dependency unit, 3 for ICU care). The study also collected measures thought to influence ICU bed occupancy including whether or not it was: at the weekend, out of hours (7 PM to 7 AM), and winter (between November and February). We expect that ICU bed availability will be correlated with these covariates, but also with their interactions. For example, the fewest beds might be available for patients assessed at night, during weekdays in the winter. We refer to the interaction of these three measures as the "timing index."

The study sample contains 13011 patients, and 10478 patients were not promptly admitted to the ICU. Of these, 2432 were later transferred to the ICU, with a mean (median) time to ICU of 22 (10) hours, with 245 patients (2.3%) transferred within 5 and 198 (1.9%) transferred within 6 hours respectively. Thus a small portion of patients, designated as controls, received ICU care just beyond the four hour window. The ICU was full at the time of 1198 (8%) assessments, but admission to the ICU can be discouraged or delayed since open beds in the ICU are often reserved for patients in surgery or not filled due to inadequate levels of ICU staffing. Next, we outline our notation based on the potential outcomes framework.

2.1 Notation

Since we use matching methods, our notation is based on a paired randomized encouragement design. After matching, there are I matched pairs for $i = 1, \ldots, I$, for $2I$ total units, and the units within matched pairs are denoted with $j \in \{1, 2\}$. Let Y_{ij} denote the outcome, an indicator of mortality at 7 (or 28) days, D_{ij} denotes the treatment actually received – prompt admission

to the ICU or not, and Z_{ij} denotes the multivalued instrument of ICU bed availability such that we observe the triplet (Y_{ij}, D_{ij}, Z_{ij}). For each individual j in pair i, let $Y_{ij}(z, d)$ be the potential outcome given the treatment assignment value $z \in \{0, 19\}$ and treatment actually received value $d \in \{0, 1\}$ and let $D_{ij}(z)$ be the potential outcome of D_{ij} given the treatment assignment value $z \in \{0, 19\}$. Observed outcomes are related to potential outcomes in the following way: $D_{ij} = D_{ij}(Z_{ij}) = Z_{ij}D_{ij}(1) + (1 - Z_{ij})D_{ij}(0)$ and $Y_{ij} = Y_{ij}(Z_{ij}, D_{ij}) = Y_{ij}(Z_{ij}, D_{ij}(Z_{ij})) = Z_{ij}Y_{ij}(1, D_{ij}(1)) + (1 - Z_{ij})Y_{ij}(0, D_{ij}(0))$.

We denote observed covariates as $\mathbf{x}_{ij}$. We form matched pairs using $\mathbf{x}_{ij}$, but we may fail to control for u_{ij} an unobserved covariate. That is, paired subjects may not be equally likely to experience a high or low value of the instrument when $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, but possibly $u_{i1} \neq u_{i2}$. Let $\mathcal{F} = \{(Y_{ij}(z, d), D_{ij}(z), \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$. Next, we write $W_{ij} = 1$ to denote $\max\{Z_{i1}, Z_{i2}\}$ for the matched pair patient with a higher number of ICU beds available, but write $W_{ij} = 0$ for the patient in the matched pair with fewer beds available. Within a matched pair, we denote the probability that a subject has the higher instrument value in the pair as $\pi_{ij} = Pr(W_{ij} = 1|\mathcal{F})$. For two subjects in the same pair i, matching ensures that observed covariates are similar ($\mathbf{x}_{i1} \approx \mathbf{x}_{i2}$) and we therefore assume that $\pi_{i1} = \pi_{i2}$. In our notation, the only random quantity is Z_{ij}, and we assume potential outcomes are fixed. This is consistent with the randomization inference framework applied to observational studies (Rosenbaum 2002, ch. 4). With a continuous instrument, such as excess beds, the ideal matched pair of subjects ik and il would have $\mathbf{x}_{ik} = \mathbf{x}_{il}$ but the difference, $Z_{ik} - Z_{il}$, will be large. That is, these units should have identical observed covariates, but one patient is strongly encouraged to receive prompt ICU care, while the other is not.

2.2 Instrumental Variable Assumptions

Identification of the causal effect of prompt ICU admission under an IV design requires a series of assumptions outlined by Angrist et al. (1996). Here, we review these assumptions and discuss their plausibility in the context of our application. First, our notation here implicitly assumes that

SUTVA holds (Rubin 1980). For SUTVA to hold, subjects outcomes must be unaffected by other patients' ICU admittance status, and the indicator D_{ij} must adequately represent all versions of the exposure. First, it seems unlikely that admitting one patient to the ICU will affect a patient on the general ward, since there are very few patients that had concurrent admissions within the same hospital. Second, we assume that while there is some variation in how patients receive care in the ICU, this variation in exposure corresponds to the same potential outcomes.

Next, we must assume that the exclusion restriction holds. Under the exclusion restriction, the outcome is only affected by the actual uptake of the treatment. Stated formally, for every individual, conditional on the exposure, the instrument has no effect on the outcome, i.e. $Y_{ij}(d,1) = Y_{ij}(d,0) = Y_{ij}(d)$ for all values of $d = 0, 1$. We judged that levels of bed occupancy meet the exclusion restriction for the following reason. The instrument is ICU bed occupancy at the time of assessment not ICU admission. ICU occupancy varies over time and according to stochastic processes, hence, even for the ICU admitted patients it is unlikely that occupancy at assessment is directly related to mortality (Gabler et al. 2013; Kahn et al. 2009). Finally, ICU staffing levels vary with the number of beds occupied such that when the ICU is full staffing levels are higher. As such, staff to patient ratios are fixed regardless of the level of ICU occupancy. Therefore, the quality of ICU care due to staffing levels should not vary with ICU occupancy.

Next, we must assume that instrument assignment is unconfounded once we condition on baseline covariates: $Y_{ij}(d), D_{ij}(z) \perp\!\!\!\perp Z_{ij}|\mathbf{x}_{ij}$. In our application, it must be the case that for similar patients in the same hospital that ICU bed availability at the time of critical care assessment varies as-if randomly. While this assumption is untestable, many authors recommend the use of a falsification test where investigators examine whether baseline covariates are balanced by instrument status (Baiocchi et al. 2014; Swanson and Hernán 2013; Davies et al. 2013; Ertefaie et al. 2017). That is, while balance in measured confounders does not provide any information about imbalances in unmeasured confounders, high levels of imbalance by instrument status casts doubt on an instrument as form of natural experiment (Rosenbaum 2010).

We calculated covariate balance for groups defined according to whether ICU bed occupancy

was above or below the median (4 beds). Table 1 reports the corresponding means and standardized differences for baseline covariates, prior to any matching. For many covariates, the imbalances are small, with just one standardized difference exceeding 0.10. In the appendix, we include plots of scaled and unscaled bias estimates with confidence intervals (Davies 2015; Jackson and Swanson 2015). These measures of imbalance do not require us to use a binary version of the instrument. Again, we find covariates tend to be balanced by the instrument. However, these results do not indicate whether there are differences in either the distribution of patients within each of the 48 hospitals or whether there are differences across the timing index.

We summarize the distribution of patients within hospital and on the timing index by reporting the total variation distance (TVD). The TVD is a distributional balance metric designed to summarize balance across many discrete categories in a single measure. A nominal covariate k with L_k levels yields an $L_k \times 2$ contingency table, and we denote the difference in counts for row ℓ of the table for covariate k as β_{kl}. As discussed in Pimentel et al. (2015) the quantity $\sum_{l=1}^{L_k} |\beta_{kl}|$ is proportional to the TVD between the two empirical probability distributions arising from the control and treated groups respectively. The TVD as expressed by $\sum_{l=1}^{L_k} |\beta_{kl}|$ may take on values ranging from 0, when the empirical distributions coincide exactly, to N which represents the total sample size when the empirical distributions share no common support. We then standardize the TVD to allow for comparisons of the TVD when sample sizes differ. Specifically, we divided by N, the total sample size, and report a scaled TVD on the interval $[0, 1]$. We calculated the TVD for the timing index and hospital. The total variation distance is 0.82, and the TVD due to differences in the distributions within hospitals is 0.55. We seek to remove this imbalance via the matching procedure we develop.

Next, we assume that the monotonicity assumption holds: for all individuals, $D_i(z') \geq D_i(z)$ for $z' > z$. The monotonicity assumption would be violated if some patients were defiers, which seems unlikely as it implies those assessing ICU admission, encouraged (discouraged) prompt transfer when there were few (many) beds available.

Finally, there must be a non-zero causal effect of Z_{ij} on D_{ij}. However, if this assumption holds,

Table 1: Balance Results For Units Above and Below the Median ICU Bed Availability at Time of Assessment.

	Less Than Median Beds Available[a] (N = 6114)	Greater Than Median Beds Available[a] (N=6897)	Std. Diff.	p-value
Age (years)	64.95	65.40	-0.03	0.15
Male	0.52	0.53	-0.00	0.95
Sepsis 0/1	0.61	0.61	-0.01	0.67
Level of Care - Level 0	0.16	0.11	0.13	0.00
Level of Care - Level 1	0.65	0.71	-0.13	0.00
Level of Care - Level 2	0.19	0.16	0.07	0.00
Level of Care - Level 3	0.01	0.01	-0.02	0.26
Rec'd Level of Care - Level 0	0.10	0.05	0.19	0.00
Rec'd Level of Care - Level 1	0.53	0.55	-0.05	0.01
Rec'd Level of Care - Level 2	0.28	0.30	-0.04	0.02
Rec'd Level of Care - Level 3	0.09	0.09	-0.00	0.91
Peri-arrest 0/1	0.05	0.05	-0.03	0.12
Weekend 0/1	0.24	0.26	-0.05	0.01
Winter 0/1	0.32	0.20	0.28	0.00
Out of Hours 0/1	0.38	0.33	0.09	0.00
Icnarc Score	15.15	15.01	0.02	0.26
News Score	6.28	6.15	0.04	0.02
Sofa Score	3.18	3.12	0.03	0.14
Level of Care Missing 0/1	0.00	0.01	-0.08	0.00
Rec'd Level of Care Missing 0/1	0.00	0.01	-0.07	0.00

Note: [a]available at time of assessment for admittance to ICU. The measure of ICU bed availability had minimum value of zero and a maximum of 19, with a mean value of 4.3 and a median value of 4. Std. Diff. – standardized difference in means.

but the effect of Z_{ij} on D_{ij} is close to zero, we may face the weak instrument problem (Bound et al. 1995). When the instrument has a weak effect on the treatment, this can result in incorrect coverage of confidence intervals (Imbens and Rosenbaum 2005; Bound et al. 1995). To assess the strength our instrument, we conducted a weak instrument test: the F-value from the weak instrument test was 30.28, which exceeds the critical value of 16.38. Thus our instrument satisfies standard tests for a weak instrument. However, Small and Rosenbaum (2008) prove that stronger instruments are more robust to bias due to departures from ignorable instrument status. Other

work has shown that stronger instruments are more resistant to bias from unobserved confounders (Baiocchi et al. 2010; Keele and Morgan 2016). Thus, we seek to both balance covariates and produce a stronger instrument. Next, we develop a matching algorithm that is tailored specifically to ensure that our design has a strong instrument and well-balanced comparison groups.

3 Near-far Matching with Refined Covariate Balance

In our application, we seek to form pairs with one patient that was assessed when there were many beds available, and the other with few, but the two individuals have similar baseline covariates. Thus we aim to create matched pairs with similar covariate values (near), but dissimilar instrument values (far). The near-far matching algorithm in Baiocchi et al. (2010) creates matched pairs that follows this template.

However, we extend standard near-far matching to account for two design features in our application. First, we want to account for variation in the selection of patients for prompt ICU admission across the 48 hospitals. In particular, assessment for ICU admission is undertaken in some hospitals by specialist critical care outreach teams, and in others by a single ICU physician. This variation in ICU admission processes implies that unmeasured covariates may be more similar within versus across hospitals. Ideally, we would compare a patient that was assessed for ICU admission at a time when the ICU had few beds available to another patient that was assessed with many beds available within the same hospital. To capture this feature, we wish to balance patients within hospitals to account for the hospital-specific ICU admission process. Second, we wish to balance the timing index to control for any correlation between the interaction of these contextual variables and demand for ICU beds. We address these requirements by introducing a new matching algorithm that adapts near-far matching to allow the specification of balance constraints. Next, we review relevant concepts.

3.1 Near-far Matching: A Review

Baiocchi et al. (2010) demonstrate how to use nonbipartite matching combined with penalties to implement the ideal IV match. In matching, penalties are used to enforce compliance with a constraint where possible, and otherwise to minimize deviation from that constraint. A near-far matching algorithm attempts to minimize distances on observables within matched pairs, subject to a penalty on instrument distance. More specifically a penalty is applied to any matched pairs that differ by more than Λ on the instrument distance as measured by $|Z_{i1} - Z_{i2}|$, where Λ is a threshold defined by the analyst. Typically a "penalty function" is applied to the within pair instrument distance. A penalty function is a continuous function that is 0 if the constraint is respected, and increases rapidly as the magnitude with which the condition is violated increases. See Rosenbaum (2010, Sec. 8.4) for a discussion of penalties in matching. Matched distances on the instrument less than Λ receive larger penalties, and are less likely to be matched. Hence, matched pairs tend to be alike on observables but different on the instrument.

In most applications increasing instrument distance leads to larger imbalances for baseline covariates. That is, as the algorithm forces pairs apart on the instrument, the matching problem becomes more restricted and some pairs formed may be insufficiently close on covariates. Baiocchi et al. (2010) remedy this problem by using "sinks" (Lu et al. 2001) to discard those observations that cannot be matched well. To eliminate e units that create suboptimal matches, e sinks are added to the data before matching, and each sink has a zero distance between each unit and an infinite distance to all other sinks. This creates a distance matrix of size $(2I + e) \times (2I + e)$. The optimal nonbipartite matching algorithm pairs e units to the e sinks to minimize the total distance between the remaining $I - e/2$ pairs. That is, by pairing a unit with a sink, the algorithm removes the e units that would form the e set of worst matches. As such, the analysts seeks a match where the penalty sufficiently strengthens the instrument, and the appropriate number of sinks yield acceptable balance. Keele and Morgan (2016) show how to use weak instrument tests to guide the choice of the penalty and sinks.

3.2 Refined Covariate Balance: A Review

We begin with a brief discussion of fine balance, which is the key component of refined covariate balance matching. For a nominal covariate k with L_k levels, with $n_{L_k} \geq 0$ treated subjects at each level, a match with κ controls is finely balanced if there are κn_{L_k} controls with level L_k of the nominal variable (Rosenbaum et al. 2007). For example, a match with fine balance on the nominal measure for hospital would produce patients paired on the IV with identical marginal distributions for hospital. The term "fine balanced" is used since a nominal covariate has been balanced at every level, but units are not exactly matched on the variable. A near-fine balance constraint returns a finely balanced match when one is feasible, and otherwise minimizes the deviation from fine balance (Yang et al. 2012). Thus with near-fine and fine balance, we place no restriction on individual matched pairs–any one treated subject can be matched to any one control, but the marginal distribution of a categorical covariate is exactly or nearly exactly the same across treated and control groups. However, even with large sample sizes, we may encounter difficulty fine balancing both hospitals and the timing index. Specifically, with 48 hospitals, 6 binary covariates and 2 nominal covariates with 4 levels, the full interaction (or joint distribution) of these variables is a nominal variable with 49,152 possible levels. Even if we balance this full interaction optimally well, many categories will likely remain imbalanced and the marginal variables may be imbalanced.

Refined covariate balance is an extension of fine or near-fine balance designed to fine balance the joint distribution of many nominal covariates (Pimentel et al. 2015). Under refined covariate balance, we define a sequence of K nested nominal covariates, which the investigator must arrange in descending order of balance prioritization. We then nest the covariates via an interaction of joint categories, such that the categories of the second covariate are all subcategories of the first covariate, and categories of the third covariate are subcategories of the previous two covariates, and so on. The algorithm then seeks to apply near fine balance to this ordering of covariate categories for each covariate in order of priority. Thus for this sequence of nested nominal variables, $v_1, \ldots, v_K$, the refined covariate balance algorithm comes as close as possible to fine

balance for v_1, and amongst all the matches that do that, it comes as close as possible to fine balance for v_2, and so on.

In our application, variable v_1 is the interaction of the 48 hospitals included in the data with the recommended level of care (i.e. v_1 includes a separate category for each distinct recommended level of care within each distinct hospital). Variable v_2 contains subcategories of variable v_1, defining whether or not patients are admitted out-of-hours, admitted on the weekend, or both. Finally, v_3 subdivides the patient groups of v_2 by current level of care. In a match with refined covariate balance, we would attempt to balance these covariates in the order outlined above. Among all matched samples that meet the refined covariate balance constraints, the algorithm minimizes the total covariate distance within pairs.

3.3 A Near-Far IV Match with Refined Covariate Balance

We develop a new matching algorithm that combines near-far matching with refined covariate balance. To do this, we depart from the original formulation of near-far matching for IV applications in two ways. First, we apply a bipartite rather than a nonbipartite match to the data. To apply a bipartite match, we applied the matching algorithm to groups defined by a binary indicator that is 1 if there are fewer than the median number of beds available in the ICU, and 0 if there are more than the median number of beds available. Next, we implemented penalties using a reverse caliper. Typically caliper matching is used to avoid poor matches by imposing a tolerance on the maximum distance between matched pairs (Cochran and Rubin 1973). For two subjects i and j, let P_i and P_j be a score on a distance metric such as the propensity score. Under a caliper, a match for subject i is selected only if $||P_i - P_j|| < \Lambda$, where Λ is a pre-specified tolerance. We reverse the concept of a caliper and only select a match for subject i if $||P_i - P_j|| > \Lambda$, where Λ remains a pre-specified tolerance. We apply the reverse caliper to the original multi-valued measure of the instrument. Applying a reverse caliper in conjunction with bipartite matching pairs units that are similar on observables, but we penalize the match unless the within pair distance on the instrument satifies the reverse caliper. Thus we build a distance matrix applying a reverse

caliper to the standard deviation of the instrument, and penalizing matches that are too close on the instrument. We then match with the refined covariate balance algorithm.

The shift to a bipartite matching enables us to use the fast, scalable balance constraint framework of Pimentel et al. (2015) rather than the much more computationally restrictive approach of Zubizarreta et al. (2013). Using bipartite matching produces matched data with the same structure as produced by nonbipartite matching: matched pairs that tend to be dissimilar on the instrument. Using a bipartite match also does not affect how we estimate the IV effect. Just as if we had applied nonbipartite matching, we calculate the average discrepancy in ICU admission status across matched pairs. See Yang et al. (2014) for a similar use of a bipartite algorithm for a near-far match.

The use of refined covariate balance does not change the basic tension between balance and instrument strength. As we strengthen the instrument, balance will tend to be worse. To increase instrument strength and balance covariates, we have to remove observations. We do so using optimal subset matching (Rosenbaum 2012), which aims to find the largest set of matched pairs such that the average matched distance within pair does not exceed a certain threshold $\widetilde{\delta}$. The parameter $\widetilde{\delta}$ can also be viewed as a penalty, describing the cost of excluding a treated individual from the match. As the value for $\widetilde{\delta}$ is decreased, more units will be excluded so that only the closest pairs will be retained in the match. While Rosenbaum (2012) considered matching without balance constraints, for a given value of $\widetilde{\delta}$ and set of refined balance constraints, our algorithm guarantees that the match produced has optimal refined balance among matches with the same number of units excluded (Pimentel and Kelz 2017). The $\widetilde{\delta}$ parameter serves a similar role to the number of sinks in the original near-far matching algorithm in that it can be used to improve balance at the cost of excluding patients from the match. Under refined covariate balance smaller values of $\widetilde{\delta}$ will also produce smaller deviations from perfectly fine balanced marginal distributions.

3.4 A General Procedure

We now give a formal specification of the algorithm. For any given set of match parameters, $(\Lambda, \widetilde{\delta})$, the match reduces to a special case of a canonical problem in mathematical optimization known as a minimum-cost network flow. In this framework, the match is represented as a graph of connected nodes. Each unit is represented as a node and distances between treated and control nodes are represented with graph edges. Each edge represents a possible match and has an associated decision variable, usually binary, referred to as a "flow." A distance metric such as the Mahalanobis distance is used to represent the distance or flow between treated and control nodes. Each edge has an associated cost per unit of flow (represented by the distance metric), and the goal of the problem is to send a certain amount of flow, or equivalently to select a certain number of matches between the treated nodes and control nodes while paying a minimal total cost. Computationally efficient algorithms exist for solving these problems in general (Bertsekas 1998). We formulate a near-far match with refined covariate balance in this framework by creating an augmented graph for the match which includes additional nodes and edges that enforce balance constraints and allow units to be excluded. We then use existing optimization methods for solving network flow problems.

We first provide an informal overview of the algorithm, and then we present the algorithm in detail.To summarize the algorithm:

1. Split the study population into two groups, using the median of the instrument as a threshold.

2. Impose calipers and exact matching constraints to form an allowable set of pairs.

3. Refine the allowable set of pairs using a reverse caliper on the instrument.

4. Construct a network flow optimization problem to solve the match, based around a bipartite graph connecting our two comparison groups and incorporating refined balance constraints and an exclusion penalty.

5. Compute the optimal flow solution and form the matched data.

We first describe the construction of this network for fixed values of Λ and $\widetilde{\delta}$, then we briefly discuss a strategy for iterating over values of these tuning parameters. Note that the notation in this section, which generally follows the notation of Pimentel et al. (2015), differs slightly from the notation given earlier in order to focus on construction of the matched sample from the raw data, rather than description of the resulting matched sample. Take the following quantities as given:

- A study population or sample of n individuals $\mathcal{S} = \{\alpha_1, \ldots, \alpha_n\}$
- An instrumental variable $Z : \mathcal{S} \longrightarrow \mathbb{R}$.
- A list of discrete variables $\nu_k : \mathcal{S} \longrightarrow \mathbb{Z}$ with $k = 1, \ldots, K$ in decreasing order of priority for balance, whose categories grow progressively finer with increasing k, so that $\nu_{k+1}(x) = \nu_{k+1}(y)$ implies $\nu_k(x) = \nu(y)$ but the converse may not hold.
- Parameters $\widetilde{\delta}$ and Λ.

Define an indicator

$$W_i = \begin{cases} 1 & \text{if} \quad Z(\alpha_i) < \text{median}(Z(\alpha_1), \ldots, Z(\alpha_n)) \\ 0 & \text{otherwise} \end{cases}$$

We use this binary version of the instrument used to structure the match as bipartite – consisting of two defined groups. We relabel the observations with $W_i = 1$ by $\tau_1, \ldots, \tau_T$ and the observations with $W_i = 0$ by $\kappa_1, \ldots, \kappa_C$ (so that $T + C = n$). Next, we write this bipartite match as a graph. Individuals τ_i form one of the groups in the bipartite graph, the high-IV group $\mathcal{T}$, while individuals κ_j form the low-IV group $\mathcal{C}$. Let $\mathcal{A} \subset \mathcal{T} \times \mathcal{C}$ represent the set of allowable pairings (τ_i, κ_j). Set $\mathcal{A}$ may be the full outer product $\mathcal{T} \times \mathcal{C}$ but more generally it may be a strict subset forbidding pairings that violate some condition; for instance, in exact matching on a nominal covariate $\mathcal{A}$ would contain only pairs with identical values of the nominal covariate, and in caliper matching

pairs would only be allowed between individuals whose values on a particular continuous covariate differ by no more than a fixed amount. In contrast to previous approaches, we refine set $\mathcal{A}$ as follows:

$$\mathcal{A}^* = \mathcal{A} \cap \{(\tau_i, \kappa_j) : |Z(\tau_i) - Z(\kappa_j)| > \Lambda\}$$

This ensures our match obeys a reverse caliper of size Λ on the instrument. For every $(\tau_i, \kappa_j) \in \mathcal{A}*$, label the covariate distance between τ_i and κ_j by c_{ij}.

Now consider the list of discrete variables $\nu_1, \ldots, \nu_K$. Label the categories of each variable ν_k by $\lambda_{k1}, \ldots, \lambda_{kL_k}$. Since the variables are nested and grow increasingly fine, each category $\lambda_{k\ell}$ with $k > 1$ has a "parent" category in the previous variable ν_{k-1}; formally, for any $k > 1$ and $\ell \in \{1, \ldots, L_k\}$, there is a single value $\ell' \in \{1, \ldots, L_{k-1}\}$ such that for any $s \in \mathcal{S}$ with $\nu_k(s) = \lambda_{k\ell}$, $\nu_{k-1}(s) = \lambda_{(k-1)\ell'}$ also. We will use the notation par$(\lambda_{k\ell})$ to represent this parent category $\lambda_{(k-1)\ell}$ in the next-most-coarse discrete variable; the parent is well-defined for $k = 2, \ldots, K$ and $j = 1, \ldots, L_k$ due to the nested structure of the covariates $\nu_1, \ldots, \nu_K$.

We can now write the match as a minimum-cost network flow problem. To do so, we define a set of nodes $\mathcal{N}$ — one for each high-IV unit and each low-IV unit in our study, plus three nodes for each category $\lambda_{k\ell}$ of our discrete variables, and one sink — and a set of directed edges $\mathcal{E}$ connecting them to one another. The treated and control units connect to one another via the bipartite graph implied by the set $\mathcal{A}^*$; the control units then connect to nodes associated with their respective categories in variable ν_K, which in turn connect to sets of nodes for their parent category in variable ν_{K-1}, and so on until nodes associated with categories of ν_1 connect to the sink node ω. In addition, we add an edge from each treated unit direct to the sink (to allow this treated unit to be discard if necessary). Formally, the graph structure is described as follows.

$$\mathcal{N} = \{\tau_1, \ldots, \tau_T, \kappa_1, \ldots, \kappa_C, \omega\} \cup \bigcup_{k=1}^{K} \{\lambda_{k1}, \lambda'_{k1}, \lambda''_{k1}, \lambda_{k2}, \lambda'_{k2}, \lambda''_{k2}, \ldots, \lambda_{kL_k}, \lambda'_{kL_k}, \lambda''_{kL_k}\}$$

$$\mathcal{E} = \mathcal{A}^* \cup \bigcup_{k=1}^{K} \bigcup_{j=1}^{L_k} \{(\lambda_{kj}, \lambda'_{kj}), \lambda_{kj}, \lambda''_{kj}), (\lambda'_{kj}, \lambda''_{kj})\} \cup \bigcup_{k=2}^{K} \bigcup_{j=1}^{L_k} \{(\lambda''_{kj}, \mathsf{par}(\lambda_{kj}))\} \cup$$
$$\bigcup_{i=1}^{C} \{(\kappa_i, \nu_K(\tau_i))\} \cup \bigcup_{j=1}^{L_1} \{(\lambda''_{1j}, \omega)\} \cup \bigcup_{i=1}^{T} \{(\tau_i, \omega)\}$$

To complete the formulation of the network flow problem, costs and capacities must be set for the flow across each edge that connects two nodes, and an amount of flow must be supplied to each node (positive amounts meaning that flow is produced there and negative amounts meaning flow is received). Assign supplies of 1 to nodes τ_i, a demand of T (or supply of $-T$) to the sink node ω, and supply of zero to all other nodes. Assign capacities of 1 to edges of form (τ_i, κ_j), $(\kappa_j, \nu_K(\kappa_j))$, and (τ_i, ω), capacities of $|\{\tau_i : \nu_k(\tau_i) = \lambda_{kj}\}|$ to edges $(\lambda_{kj}, \lambda''_{kj})$, and infinite capacity to all other edges. Additionally, assign costs c_{ij} to edges (τ_i, κ_j), cost $\widetilde{\delta}$ to edges (τ_i, ω), costs Υ^{K-k+1} to edges $(\lambda_{kj}, \lambda'_{kj})$ where $\Upsilon >> \max c_{ij}$, and cost 0 to all other edges. Note that this network is identical to the one in Pimentel et al. (2015) except for the inclusion of the bypass edges (τ_i, ω) – although these are described in the context of the original algorithm in the appendix to Pimentel and Kelz (2017) – and the inclusion of the reverse caliper. When the minimum-cost network flow problem defined by this network is solved using a standard generalized solver, the pairings (τ_i, κ_j) with nonzero flow in the optimal solution form the optimal match.

Notice that the above algorithm assumes that Λ and $\widetilde{\delta}$ are fixed in advance. In practice, we recommend that researchers consider several solutions using a grid search over values of these parameters and select a single configuration that balances the tradeoff between sample size and instrument strength well. Larger values of Λ and smaller values of $\widetilde{\delta}$ respectively tend to reduce the sample size and improve the instrument strength. Finally, we note that balancing marginal distributions may come at some cost in terms of pairwise distances, since we are no longer focusing on the full distribution of the covariates using the pairwise distance alone. This implies that while

a match of this type is the optimal solution for this representation of the match, it is not optimal in terms of the pairwise distances. However, researchers can check for any associated increase in the pairwise distances using post-match balance checks.

4 The Match

4.1 How the matching was done

Prior to matching, we calculated the pairwise distances between the patients included in the sample. We used a rank-based Mahalanobis distance metric, which unlike the usual Mahalanobis distance, is robust to low-incidence binary variables and variables with highly skewed distributions (Rosenbaum 2010). For two covariates, the current level of care, and the recommended level of care, a small fraction of data were missing. Instead of imputing these missing values based on a model, we use a method recommended by Rosenbaum (2010). Specifically, we imputed missing values using the mean for that covariate. We then created a separate indicator for whether the value was missing and included these indicators within the matching algorithm to ensure the pattern of missingness for each variable was balanced across comparison groups.

The first match we implemented strengthened the instrument but did not include any refined covariate balance constraints. We conducted matches with several values for Λ and $\widetilde{\delta}$. We used weak instrument tests to select values of $\Lambda = 1.5$ and $\widetilde{\delta} = 1000$. This implies that the minimum discrepancy between paired individuals was 1.5 standard deviations of the instrument. We later examine whether the study conclusions are sensitive to these choices. We implemented a second match which retained the same parameters for Λ and $\widetilde{\delta}$, but also included refined covariate balance constraints. We added refined covariate balance constraints for the hospital, the timing index, and the nominal measures for existing level of care at assessment, and recommended level of care after assessment. In both matches, we exactly matched on winter, since it displayed the largest imbalance in the data.

4.2 Balance Results

Table 2 contains means and standardized differences for both matches. The standardized difference on the instrument for both matches is approximately three. As such, both matches were similar in the extent to which they strengthened the instrument. The balance on the univariate patient-level covariates is slightly better in the second match that includes refined covariate balance. Here, none of the standardized differences exceed 0.05, whereas following the first match, three of the standardized difference are 0.10 or larger.

Table 2: Covariate balance and degree of encouragement in two matched comparisons: one without refined covariate balance and one with.

	Stronger Instrument W/o Refined Covariate Balance 4596 Pairs of Patients			Stronger Instrument With Refined Covariate Balance 2048 Pairs of Patients		
	Few Beds Available[a] Mean	More Beds Available[a] Mean	Std. Diff.	Few Beds Available[a] Mean	More Beds Available[a] Mean	Std. Diff.
Available ICU Beds	1.68	7.64	2.91	1.56	7.05	3.07
Age (years)	65.00	65.23	0.01	64.80	65.94	0.06
Male	0.53	0.53	0.01	0.54	0.54	0.00
Sepsis 0/1	0.63	0.62	0.00	0.62	0.62	0.00
Level of Care - Level 0	0.13	0.11	0.07	0.10	0.10	0.01
Level of Care - Level 1	0.69	0.71	0.04	0.70	0.70	0.01
Level of Care - Level 2	0.17	0.16	0.02	0.18	0.18	0.01
Level of Care - Level 3	0.01	0.01	0.05	0.01	0.01	0.00
Rec'd Level of Care - Level 0	0.08	0.04	0.17	0.04	0.04	0.01
Rec'd Level of Care - Level 1	0.55	0.55	0.00	0.53	0.52	0.02
Rec'd Level of Care - Level 2	0.28	0.30	0.05	0.32	0.32	0.01
Rec'd Level of Care - Level 3	0.08	0.09	0.04	0.11	0.11	0.02
Peri-arrest 0/1	0.04	0.05	0.07	0.05	0.05	0.02
Weekend 0/1	0.23	0.26	0.06	0.24	0.25	0.02
Winter 0/1	0.21	0.21	0.00	0.27	0.27	0.00
Out of Hours 0/1	0.36	0.34	0.05	0.36	0.34	0.05
Icnarc Score	15.23	15.07	0.02	15.59	15.57	0.00
News Score	6.28	6.18	0.03	6.42	6.28	0.05
Sofa Score	3.16	3.14	0.01	3.31	3.27	0.02
Level of Care Missing 0/1	0.00	0.01	0.10	0.00	0.00	0.04
Rec'd Level of Care Missing 0/1	0.00	0.01	0.11	0.00	0.00	0.01
Total Variation Distance		0.71			0.09	
Total Variation Distance Hospital Only		0.66			0.05	

Note: [a]at time of assessment for admittance to ICU. The measure of ICU bed availability had minimum value of zero and a maximum of 19, with a mean value of 4.3 and a median value of 4. Std. Diff. – absolute standardized difference in means.

We summarize the effect of applying the refined covariate balance constraints on the distribution of patients within hospital by reporting the TVD. After matching, we standardize the TVD using $2I$ the total number of patients in the match. We calculated the TVD for each variable separately and then summed these values. These statistics are reported in the bottom panel of Table 2. The matching with refined covariate balance led to much smaller TVD (0.09 vs 0.71) in particular across hospitals (0.05 versus 0.66). Thus the use of refined covariate balance reduces the total variation distance far more than just using near-far matching. This comes at the cost of having to discard a large number of patients. The appendix contains a comparison between patients in each matched sample and the patients discarded by the match.

For the two matches in Table 2, we see that we can maintain a stronger IV and produce better fine balance by adding refined covariate balancing constraints. However, the relative roles of the refined covariate balance constraints and the sample trimming are unclear. We therefore undertook two further matches. For the first match, we targeted the sample size of the match in Table 2 without refined covariate balance, but then added refined covariate balance constraints. For the second match, we targeted the sample size of the match with refined covariate balance in Table 2, but removed the refined balance constraints. These additional matches allow us to isolate the role of refined covariate balance from the use of $\widetilde{\delta}$. We also report for each match the average distance between matched pairs using the Mahalanobis distance. The results are in Table 3.

Table 3: Total variation distance and average pair distances in the matches without and with refined covariate balance constraints for fixed sample sizes.

	Without Refined Covariate Balance 4596 Pairs of Patients	With Refined Covariate Balance 4596 Pairs of Patients
Total Variation Distance	0.71	0.60
Total Variation Distance Hospital Only	0.66	0.55
Average Pair Distance	22.45	22.53
	Without Refined Covariate Balance 2040 Pairs of Patients	**With Refined Covariate Balance 2048 Pairs of Patients**
Total Variation Distance	0.68	0.09
Total Variation Distance Hospital Only	0.68	0.05
Average Pair Distance	20.33	22.59

First, we observe that, even without trimming the sample, the addition of refined covariate balance does improve balance as the TVD declines from 0.71 to 0.60 when the sample size is fixed at 4,596 matched pairs. Moreover, trimming the sample size alone does little to improve balance on the marginal distributions. In a match that trims the data but does not add refined covariate balance, the TVD declines little from 0.71 to 0.68. The combination of trimming and refined covariate balance clearly produces the lowest value for the TVD (0.09). Second, matches that do not involve refined balance constraints have lower average pair distances than those that incorporate them. This is expected; prioritizing marginal balance over close pairing results in sub-optimal pair distances. However, the relative increase in average pair distance associated with the addition of a marginal balance constraint is small compared to the relative decrease in marginal imbalance. Even with the smaller sample sizes, where the increase in average pair distance arising from the inclusion of balance constraints is over 10% (from 20.33 to 22.59), the percentage reduction in balance exceeds 90%. The change in closeness of individual pairs is even less consequential when measured on any one of the many variables that are used to create the summary Mahalanobis distance. Thus although marginal balance comes at the cost of increased pair distances, the magnitude of that increase is negligible; especially in light of the balance statistics reported in Table 2. However, the gains in terms of the TVD are substantial.

5 Analyzing the Matched IV Design

Under an IV design, the causal estimand can be charactized using the generalized effect ratio from Baiocchi et al. (2010). Formally, the generalized effect ratio is:

$$\lambda = \frac{\sum_{i=1}^{I} \sum_{j=1}^{2} [Y_{ij}(1, D_{ij}(1)) - Y_{ij}(0, D_{ij}(0))]}{\sum_{i=1}^{I} \sum_{j=1}^{2} [D_{ij}(1) - D_{ij}(0)]} \tag{1}$$

The effect ratio measures the relative magnitude of two treatment effects. The first treatment in the numerator is the effect of ICU bed availability on mortality. The second treatment effect, in

the denominator, is the effect of ICU bed availability on prompt ICU admission. In a randomized trial, these two treatment effects are estimable, however, further assumptions are needed to link these two effects. Under the exclusion restriction and monotonicity, we can interpret λ as the average decrease in mortality caused by prompt ICU care, amongst those patients who only received prompt ICU care because there were beds available at the time of assessment on the general ward (Angrist et al. 1996).

To estimate λ, we use the following estimator

$$\hat{\lambda} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{2} E(Y_{ij}|W_{ij} = 1) - E(Y_{ij}|W_{ij} = 0)}{\sum_{i=1}^{I} \sum_{j=1}^{2} E(D_{ij}|W_{ij} = 1) - E(D_{ij}|W_{ij} = 0))} \tag{2}$$

where $W_{ij} = 1$ for the matched pair unit with a higher value of the instrument and is 0 otherwise. Equation 2 is often referred to as the "Wald" estimator (Wald 1940; Hernán and Robins 2006; Baiocchi et al. 2014) and is equivalent to estimation via two-stage least squares (Baiocchi et al. 2014). We test Fisher's sharp null hypothesis of no treatment effect, where randomization forms the "reasoned basis for inference" (Fisher 1935). The sharp null asserts that $H_0 : Y_{ij}(1, D_{ij}(1)) = Y_{ij}(0, D_{ij}(0))$ for all i and j, which implies that changes in the number of excess ICU beds has no impact on mortality. If the exclusion restriction holds, this sharp null hypothesis also implies that prompt ICU care has no impact on mortality. We can test $\lambda = 0$ using McNemar's test. Alternatively, finite sample approximations to the exact test are available. Given the relatively large sample sizes in our application, we use the finite sample approximation. We test this null hypothesis using a test statistic derived in Baiocchi et al. (2010). This test statistic is comprised of two terms, the first being:

$$T(\lambda_0) = \frac{1}{I} \sum_{i=1}^{I} \left\{ \sum_{j=1}^{2} Z_{ij}(Y_{ij} - \lambda_0 D_{ij}) - \sum_{j=1}^{2} (1 - Z_{ij})(Y_{ij} - \lambda_0 D_{ij}) \right\} = \frac{1}{I} \sum_{i=1}^{I} V_i(\lambda_0)$$

The second term has the following form:

$$S^2(\lambda_0) = \frac{1}{I(I-1)} \sum_{i=1}^{I} \{V_i(\lambda_0) - T(\lambda_0)\}^2$$

We can test $H_0 : \lambda = 0$ by comparing $T(\lambda_0)/S(\lambda_0)$ to a standard Normal cumulative distribution. See Kang et al. (2018) for a review of exact inferential methods for for IV methods.

6 IV Estimates of Prompt ICU Care

In Table 4, we report the estimated effect of prompt ICU care on 7- and 28-day mortality for the first two matches with and without refined covariate balance. For the first match, the point estimate for 7-day mortality is -0.03 [95% confidence interval -0.210, 0.144], and for the match with refined covariate balance constraints, the point estimate is -0.25 [-0.642, 0.078]. For the 28-day mortality endpoint, the corresponding point estimates [95% confidence interval] are -0.12 [-0.342, 0.088] for the match without, and -0.19 [-0.638, 0.216], for the match with, refined covariate balance constraints. While our estimates our imprecise as indicated by the wide confidence intervals, they agree with estimates from other IV studies of ICU care (Kc and Terwiesch 2012; Shmueli et al. 2004; Valley et al. 2015).

What might account for the differences in the outcome estimates by match type? The primary difference between the two matches was that refined covariate balancing allowed us to nearly exactly balance patients within hospitals. As we outlined above, hospitals have specific ICU admission procedures and balancing the marginal distribution of hospitals better allows us to control for hospital specific aspects of the instrument assignment mechanism. Thus if we assume that balancing patients within hospitals is a critical component of the outcome model, we would favor the estimates under refined covariate balancing.

We consider whether these estimates are sensitive to the choice of reverse caliper for instrument distance within matched pairs. We re-estimated the treatment effects for matches with the IV caliper set to 1.0 and 2.0, which implies a somewhat weaker, or stronger instrument, compared to the estimates in Table 4where the caliper was set to 1.5. When the match with the refined balance constraint is repeated with a caliper of 2.0, the estimated effects of prompt ICU admission are -0.02 (7-day mortality) and -0.21 (28-day mortality), whereas with a caliper of 1.0,

Table 4: Estimated Effect of Prompt ICU Admission on 7 and 28 Day Mortality.

Strong IV Without Refined Covariate Balance			
	Point Estimate	95% CI	p-value
7 Day Mortality	-0.031	[-0.210, 0.144]	0.73
28 Day Mortality	-0.122	[-0.342, 0.088]	0.256
Strong IV With Refined Covariate Balance			
	Point Estimate	95% CI	p-value
7 Day Mortality	-0.252	[-0.642, 0.078]	0.132
28 Day Mortality	-0.189	[-0.638, 0.216]	0.351

the corresponding estimates are 0.03, and 0.13. For the match which did not include the refined covariate balance constraints, the corresponding estimates are -0.04 (7-day mortality) and -0.09 28-day mortality, with the caliper set to 2.0, and -0.09, and -0.14 with the caliper set to 1.0. In all cases, the 95% confidence intervals include zero.

7 Discussion

This paper develops a new near-far matching algorithm that can improve IV strength while balancing many nominal categories. Our extension is motivated by the fact that the treatment selection process may differ on unobserved characteristics at the hospital-level. We addressed this problem by combining near-far matching with refined covariate balance constraints, to ensure that patients were near finely balanced on the hospital of admission and other key covariates. Our contribution is part of a wider literature that focuses on the design of IV methods in clinical settings where unobserved confounding is present, but randomized trials are infeasible (Baiocchi et al. 2012). One advantage of this design-based approach to IV methods is that once the matching is complete, treatment effects can be estimated with simple, transparent methods.

This study also contributes to the literature on the effectiveness of ICU admission. Critical care is one of many settings where randomization is not possible due to ethical guidelines. Moreover, confounding by indication likely threatens studies that rely on the assumption that all potential confounders have been observed (Simchen et al. 2007). While our estimates are not precisely

estimated, the results show that prompt ICU care leads to an average reduction in 7- and 28-day mortality. Thus are results are consistent with work that has used IV designs to estimate the effect of ICU care (Hu et al. 2018; Shmueli et al. 2004; Harris et al. 2018; Valley et al. 2015).

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Baiocchi, M., Cheng, J., and Small, D. S. (2014), "Instrumental variable methods for causal inference," *Statistics in medicine*, 33, 2297–2340.

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010), "Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants," *Journal of the American Statistical Association*, 105, 1285–1296.

Baiocchi, M., Small, D. S., Yang, L., Polsky, D., and Groeneveld, P. W. (2012), "Near/far matching: a study design approach to instrumental variables," *Health Services and Outcomes Research Methodology*, 12, 237–253.

Bertsekas, D. P. (1998), *Network optimization: continuous and discrete models*, Belmont, MA: Athena Scientific.

Bound, J., Jaeger, D., and Baker, R. (1995), "Problems with Intrustmental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable Is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Chalfin, D. B., Trzeciak, S., Likourezos, A., Baumann, B. M., Dellinger, R. P., study group, D.-E., et al. (2007), "Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit*," *Critical care medicine*, 35, 1477–1483.

Cochran, W. G. and Rubin, D. B. (1973), "Controlling Bias in Observational Studies," *Sankyha-Indian Journal of Statistics, Series A*, 35, 417–446.

Davies, N. M. (2015), "Commentary: an even clearer portrait of bias in observational studies?" *Epidemiology (Cambridge, Mass.)*, 26, 505.

Davies, N. M., Smith, G. D., Windmeijer, F., and Martin, R. M. (2013), "Issues in the reporting and conduct of instrumental variable studies: a systematic review," *Epidemiology*, 24, 363–369.

Ertefaie, A., Small, D. S., Flory, J. H., and Hennessy, S. (2017), "A tutorial on the use of instrumental variables in pharmacoepidemiology," *Pharmacoepidemiology and Drug Safety*, 26, 357–367.

Fisher, R. A. (1935), *The Design of Experiments*, London: Oliver and Boyd.

Gabler, N., Ratcliffe, S., Wagner, J., Asch, D., Rubenfeld, G., Angus, D., and Halpern, S. (2013), "Mortality among patients admitted to strained intensive care units." *Am J Respir Crit Care Med*, 188, 800–806.

Harris, S., Singer, M., C, C. S., Grieve, R., Harrison, D., and Rowan, K. (2018), "Impact on mortality of prompt admission to critical care for deteriorating ward patients: an instrumental variable analysis using critical care bed strain," *Intensive Care Medicine*, in press.

Harris, S., Singer, M., Rowan, K., and Sanderson, C. (2015), "Delay to admission to critical care and mortality among deteriorating ward patients in UK hospitals: a multicentre, prospective, observational cohort study," *The Lancet*, 385, S40.

Hernán, M. A. and Robins, J. M. (2006), "Instruments for Causal Inference: An Epidemiologists Dream," *Epidemiology*, 17, 360–372.

Hu, W., Chan, C. W., Zubizarreta, J. R., and Escobar, G. J. (2018), "An examination of early transfers to the ICU based on a physiologic risk score," *Manufacturing & Service Operations Management*.

Imbens, G. W. and Rosenbaum, P. (2005), "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education," *Journal of The Royal Statistical Society Series A*, 168, 109–126.

Jackson, J. W. and Swanson, S. A. (2015), "Toward a clearer portrayal of confounding bias in instrumental variable applications," *Epidemiology*, 26, 498.

Kahn, J., Ten Have, T., and Iwashyna, T. (2009), "The relationship between hospital volume and mortality in mechanical ventilation: an instrumental variable analysis." *Health Services Research*, 44, 862–879.

Kang, H., Peck, L., and Keele, L. (2018), "Inference for Instrumental Variables: A Randomization Inference Approach," *Journal of The Royal Statistical Society, Series A*, In press.

Kc, D. S. and Terwiesch, C. (2012), "An econometric analysis of patient flows in the cardiac intensive care unit," *Manufacturing & Service Operations Management*, 14, 50–65.

Keele, L. J. and Morgan, J. (2016), "How Strong is Strong Enough? Strengthening Instruments Through Matching and Weak Instrument Tests," *Annals of Applied Statistics*, 10, 1086–1106.

Lu, B., Zutto, E., Hornik, R., and Rosenbaum, P. R. (2001), "Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse," *Journal of the American Statistical Association*, 96, 1245–1253.

Pimentel, S. D. and Kelz, R. (2017), "Optimal Tradeoffs in Matching Designs for Observational Studies," Unpublished Manuscript.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, Sparse Optimal Matching with Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Pirracchio, R., Sprung, C., Payen, D., and Chevret, S. (2011), "Benefits of ICU admission in critically ill patients: Whether instrumental variable methods or propensity scores should be used," *BMC medical research methodology*, 11, 1.

Renaud, B., Santin, A., Coma, E., Camus, N., Van Pelt, D., Hayon, J., Gurgui, M., Roupie, E., Hervé, J., Fine, M. J., et al. (2009), "Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia*," *Critical care medicine*, 37, 2867–2874.

Rosenbaum, P. R. (2002), *Observational Studies*, New York, NY: Springer, 2nd ed.

— (2010), *Design of Observational Studies*, New York: Springer-Verlag.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Mimimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer," *Journal of the American Statistical Association*, 102, 75–83.

Rubin, D. B. (1980), "Bias reduction using Mahalanobis-metric matching," *Biometrics*, 36, 293–298.

Shmueli, A., Baras, M., and Sprung, C. L. (2004), "The effect of intensive care on in-hospital survival," *Health Services and Outcomes Research Methodology*, 5, 163–174.

Simchen, E., Sprung, C. L., Galai, N., Zitser-Gurevich, Y., Bar-Lavi, Y., Levi, L., Zveibil, F., Mandel, M., Mnatzaganian, G., Goldschmidt, N., Ekka-Zohar, A., and Weiss-Salz, I. (2007), "Survival of critically ill patients hospitalized in and out of intensive care." *Crit Care Med*, 35, 449–457.

Small, D. and Rosenbaum, P. R. (2008), "War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases," *Journal of the American Statistical Association*, 103, 924–933.

Swanson, S. A. and Hernán, M. A. (2013), "Commentary: how to report instrumental variable analyses (suggestions welcome)," *Epidemiology*, 24, 370–374.

The Intensive Care Society (2013), *Core Standards for Itensive Care*, London, UK: Intensive Care Society, 1st ed.

Valley, T., Sjoding, M., Ryan, A., Iwashyna, T., and Cooke, C. (2015), "Association of intensive care unit admission with mortality among older patients with pneumonia," *JAMA*, 314, 1272–1279.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300.

Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes," *Biometrics*, 68, 628–636.

Yang, F., Zubizaretta, J., Small, D. S., Lorch, S., and Rosenbaum, P. (2014), "Dissonant Conclusions When Testing the Validity of an Instrumental Variable," *The American Statistician*, 68, 253–263.

Zubizarreta, J. R., Small, D. S., Goyal, N. K., Lorch, S., and Rosenbaum, P. R. (2013), "Stronger Instruments via Interger Programming in an Observational Study of Late Preterm Birth Outcomes," *Annals of Applied Statistics*, 7, 25–50.

Appendices

A.1 Bias Plots

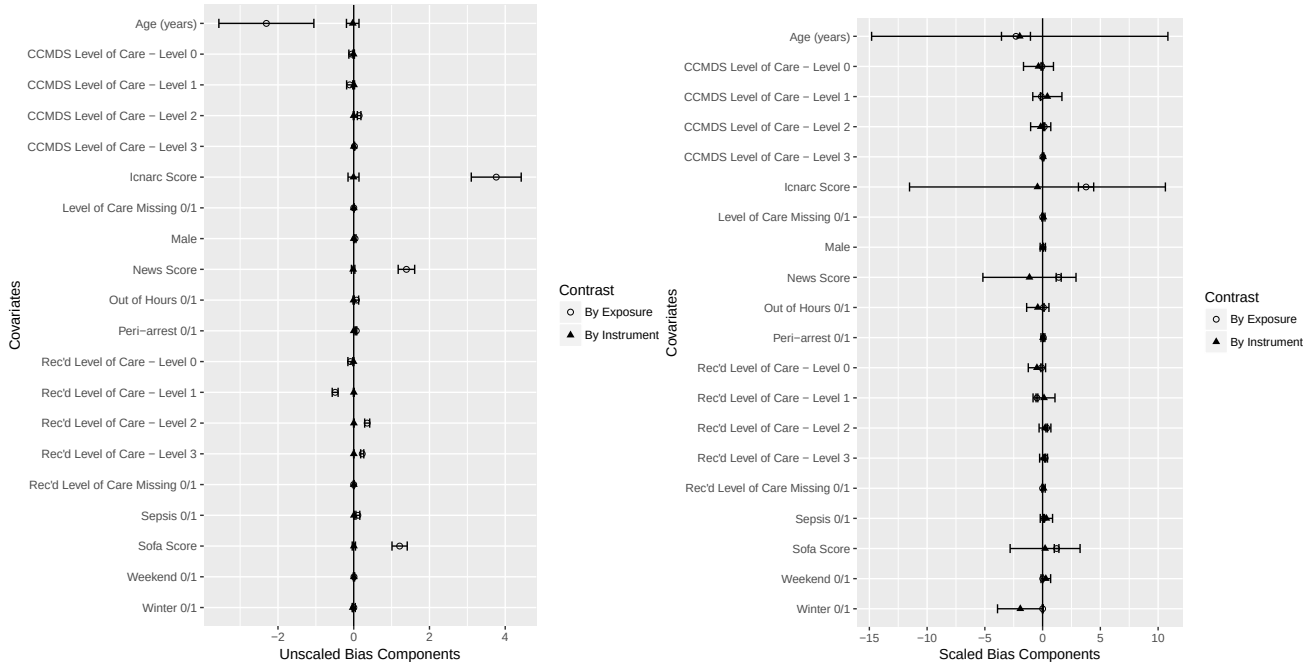

Figure 1: Bias Plots for Covariates

A.2 Summary Statistics for Patients Excluded from the Matches

Table 5: Balance Results for Observations Excluded from the Strong IV match without RC Balance

	Included in Match			Excluded From Match		
	Few Beds Available Mean	More Beds Available Mean	Std. Diff.	Few Beds Available Mean	More Beds Available Mean	Std. Diff.
Available ICU Beds	1.68	7.64	-2.91	1.94	4.25	-3.40
Age	65.00	65.23	-0.01	64.81	65.76	-0.05
Male	0.53	0.53	0.01	0.50	0.52	-0.04
Sepsis 0/1	0.63	0.62	0.00	0.56	0.59	-0.06
Level of Care	1.05	1.08	-0.05	1.03	1.05	-0.03
Rec'd Level of Care	1.37	1.45	-0.12	1.38	1.42	-0.05
Peri-arrest 0/1	0.04	0.05	-0.07	0.08	0.06	0.09
Weekend	0.23	0.26	-0.06	0.26	0.26	-0.01
Winter	0.21	0.21	0.00	0.67	0.19	1.10
Out of Hours	0.36	0.34	0.05	0.43	0.32	0.22
Icnarc Score	15.23	15.07	0.02	14.94	14.88	0.01
News Score	6.28	6.18	0.03	6.28	6.09	0.06
Sofa Score	3.16	3.14	0.01	3.23	3.09	0.06
Level of Care Missing	0.00	0.01	-0.10	0.01	0.01	-0.03
Rec'd Level of Care Missing 0/1	0.00	0.01	-0.10	0.01	0.01	0.02

Table 6: Balance Results for Observations Excluded from the Strong IV match with RC Balance

	Included in Match			Excluded From Match		
	Few Beds Available Mean	More Beds Available Mean	Std. Diff.	Few Beds Available Mean	More Beds Available Mean	Std. Diff.
Available ICU Beds	1.56	7.05	-3.07	1.84	6.28	-2.07
Age	64.80	65.94	-0.06	65.03	65.18	-0.01
Male	0.54	0.54	-0.00	0.52	0.52	-0.00
Sepsis 0/1	0.62	0.62	-0.00	0.60	0.61	-0.01
Level of Care	1.10	1.11	-0.00	1.02	1.05	-0.06
Rec'd Level of Care	1.49	1.51	-0.03	1.31	1.41	-0.14
Peri-arrest 0/1	0.05	0.05	-0.02	0.05	0.05	-0.03
Weekend	0.24	0.25	-0.02	0.23	0.26	-0.06
Winter	0.27	0.27	0.00	0.35	0.17	0.41
Out of Hours	0.36	0.34	0.05	0.38	0.33	0.11
Icnarc Score	15.59	15.57	0.00	14.93	14.77	0.02
News Score	6.42	6.28	0.05	6.20	6.10	0.03
Sofa Score	3.31	3.27	0.02	3.11	3.06	0.02
Level of Care Missing	0.00	0.00	-0.04	0.00	0.01	-0.09
Rec'd Level of Care Missing 0/1	0.00	0.00	-0.04	0.01	0.01	-0.08